A very important part of statistics is the study of the sorts of conclusions we can make about an entire population on the basis of a relatively small sample. The field of inferential statistics enables you to make educated guesses about the numerical characteristics of large groups.

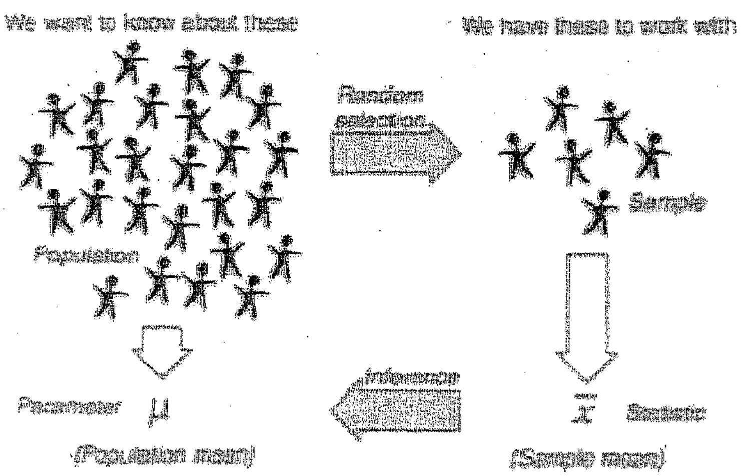
A population is a group that have something in common. Population is the term statisticians use to describe a large set or collection of items that have something in common.

In medicine, there are several populations of interest (e.g. intensive care unit patients, patients with acute respiratory distress syndrome, or patients who receive renal replacement therapy). It is seldom possible to obtain information from every individual in these populations, however, and attention is more commonly restricted to samples drawn from them.

A sample is a subset of the population, selected to representative of the larger population. It is essential that any sample is as representative as possible of the population from which it is drawn.

Often, researchers want to know things about populations but don't have data for every person or thing in the population. If a company's customer service division wanted to learn whether its customers were satisfied, it would not be practical or perhaps even possible to contact every individual who purchased a product. Instead, the company might select a sample or samples of the population. A sample is a smaller group of members of a population selected to represent the population. In order to use statistics to learn things about the population, the sample must be random. A random sample is one in which every member of a population has an equal chance to be selected.

A parameter is a characteristic of a population. A statistic is a characteristic of a sample. Inferential statistics enables you to make an educated guess about a population parameter based on a statistic computed from a sample randomly drawn from that population.

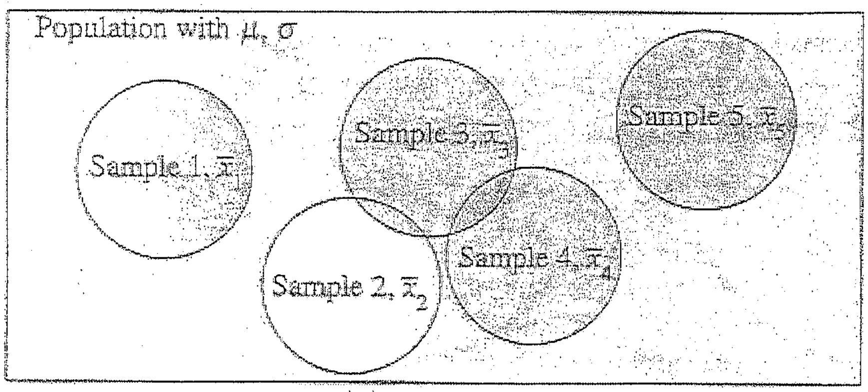


Again, the purpose of dealing with a sample is straightforward: it enables us to study a large population and to learn things about it, so that we can draw important inferences, without having to go to the trouble of collecting data from every member of the entire population.

A sampling distribution is the probability distribution of a sample statistic that is formed when samples of size n are repeatedly taken from a population. If the sample statistic is the sample mean, then the distribution is the sampling distribution of sample means.

The sampling distribution of the mean is a very important distribution. Later you will see that it is used to construct confidence intervals for the mean and for significance testing.

The Venn Diagram below represents a large population and each circle represents a sample of size n.



If you compute the mean of a sample of 10 numbers, the value you obtain will probably not equal the population mean exactly. It will be a little bit higher or a little bit lower. If you sampled sets of 10 numbers over and over again (computing the mean for each set), you would find that some sample means come much closer to the population mean than others. Some would be higher than the population mean and some would be lower and some would be spot on.

Imagine sampling 10 numbers and computing the mean over and over again, say about 1,000 times, and then constructing a relative frequency distribution of those 1,000 means. This distribution of means is a very good approximation to the sampling distribution of the mean. The sampling distribution of the mean is a theoretical distribution of the population that is approached as the number of samples increases.

Properties of Sampling Distributions of Sample Means

1. The mean of the sample means $\mu_{\bar{x}}$ is equal to the population mean μ. $\mu_{\bar{x}} = \mu$

2. The standard deviation of the sample means $\sigma_{\bar{x}}$ is equal to the population standard deviation σ divided by the square root of n. $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

The standard deviation of the sampling distribution of the sample means is called the standard error of the mean. A very important implication of this standard errorformula is that you must quadruple the sample size (4×) to achieve half (1/2) the measurement error. When designing statistical studies where cost is a factor, this may have a factor in understanding cost-benefit tradeoffs.